DOCUMENT RESUME

ED 400 317 TM 025 713

AUTHOR Ito, Kyoko; Sykes, Robert C.
TITLE A Comparison of Three Equating Approaches to A Random-Groups, Common-Forms Design.
PUB DATE Apr 96
NOTE 31p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Equated Scores; *Error of Measurement; Grade 11; High Schools; High School Students; *Statistical Bias; *Test Format
IDENTIFIERS *Calibration; *Chained Equipercentile Equating

ABSTRACT

Equating multiple test forms is frequently desired. When multiple forms are linked in a chain of equating, error tends to build up in the process. This paper compares three procedures for equating multiple forms in a common-form design where each school administered, in a spiraled fashion, only a subset of multiple forms. Data used were from a tryout administration of a grade-11 proficiency examination for which 10 forms of reading, 10 of science, and 9 of mathematics examinations were administered. The results were evaluated with regard to the amount of equating error. The three approaches differed in terms of the presence or absence of "explicit" equating of equivalent groups based on the observed ability distributions and whether calibration samples were combined across common forms or not. The approach that involved explicit equating and combined-group calibrations resulted in less bias and standard error. (Contains two numbered tables, four numbered figures, five unnumbered tables, two unnumbered figures, and nine references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
KYOKO ITO
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 400 317

A comparison of three equating approaches to a random-groups, common-forms design

Kyoko Ito

Robert C. Sykes

CTB/McGraw-Hill

This paper was presented at the Annual Meeting of the National Council on Measurement in Education, New York, 1996.

TM025713

Abstract

It is frequently desired to equate multiple test forms. When multiple forms are linked in a chain of equating, error tends to build up in the process. This paper compared three procedures for equating multiple forms in a common-form design where each school administered, in a spiraled fashion, only a subset of multiple forms. The results were evaluated with regard to the amount of equating error.

The three approaches differed in terms of (1) presence or absence of "explicit" equating of equivalent groups based on the observed ability distributions, and (2) whether calibration samples were combined across common forms, or not. The approach that involved explicit equating and combined-group calibrations resulted in less bias and standard error.

Introduction

When multiple forms are to be linked in a chain of equating, error tends to build up in the process. The accumulated drift would likely appear when the beginning and end of the chain are connected. This paper examined three procedures for equating multiple forms in a chain and evaluated them with regard to the amount of equating error.

Suppose ten forms of a test have been constructed. The forms are very similar to one another in terms of difficulty and reliability. Only three or four of the ten forms will be administered to any one school in a state, because administering all ten forms to each school may jeopardize security of the items. A set of forms will be spiraled within classrooms in a given school. Thus, the ten forms will first be split into six sets. In grouping the forms into sets, a common link between sets is necessary to make subsequent horizontal equating of the forms possible. Suppose further the following sets and linking design have been adopted:

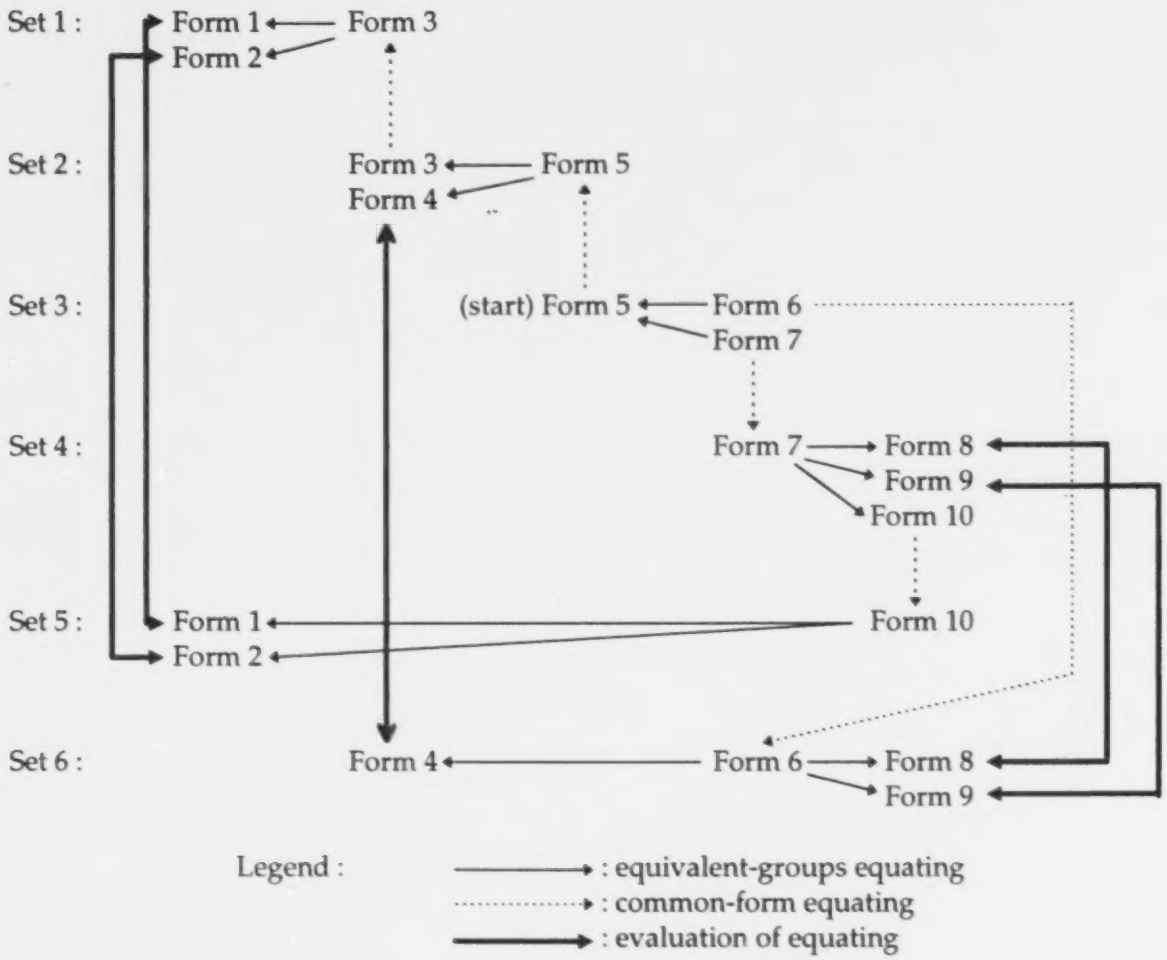
Set	# of Forms	Form
1	3	1, 2, **3**
2	3	**3**, 4, **5**
3*	3	**5**, 6, **7**
4	4	**7**, 8, 9, **10**
5	3	**10**, 1, 2
6	4	**4**, 6, 8, 9

* Beginning of the chaing of linked forms.

Except for Set 6, adjacent sets are linked by the last and first forms in the sets (bolded form numbers in the "Form" column in the table above). Set 6 connects to the beginning

set 3 through Form 6 (underlined).

The linking design is schematically depicted below:



The above design shows the chain of equatings and the kind of equating for each link: equivalent-groups equating, common-form equating, or evaluative equating. Equivalent-groups equating (thin line) is appropriate within sets, because as a result of spiraling, samples are expected to be randomly equivalent within a set of

forms. However, random equivalence of samples cannot be expected across sets of forms, due to lack of spiraling. Thus, common-form equating (broken line) is required across sets. Finally, the quality of linking must be evaluated through the equating of a form to itself (thick line).

In the above example, several equating approaches are possible for two reasons. First, spiraling may provide, within sets, samples that are truly equivalent, like a single sample, or yield samples that are not. On the one hand, if spiraling has produced truly equivalent samples, there is no need for observed-score equating of the samples. This is because calibration programs often operate under the implicit assumption that a true ability distribution has certain mean and standard deviations. Through this assumption, equivalent samples will be implicitly equated in the process of calibration. On the other hand, if random samples are not equivalent within sets, the observed ability distributions must be used to equate them explicitly. Thus, two options are available in the above illustration: explicit equating of equivalent groups based on the observed ability distributions, or no such explicit equating (i.e., only implicit equating).

Second, forms can be calibrated using "individual" samples or "combined" samples. This is because each form was administered to two samples in two sets. In the "individual-sample" calibration of forms, for example, Form 1 would be calibrated twice: first, using a sample who took Form 1 in Set 1, and second, using a sample who

took Form 1 in Set 5. In the "combined-sample" calibration, Form 1 would be calibrated using a combined sample who took the form in either Set 1 or Set 5 . Only one set of estimates would be produced per form in the "combined-sample" approach, as opposed to two sets of estimates in the "individual-sample" approach.

The paper compared the following three equating approaches:

1. Individual-sample calibration of forms and "explicit" equivalent-groups equating of observed ability distributions,
2. Individual-sample calibration of forms but no "explicit" equivalent-groups equating of observed ability distributions under the assumption that the groups are equivalent,
3. Combined-sample calibration of forms and "explicit" equivalent-groups equating of observed ability estimates.

Data

The study used data from a tryout administration of a high-school (grade 11) proficiency examination. The examination was administered in the fall of 1994 in three content areas: Reading, Mathematics, and Science. The table below shows, for each content area, the number of forms administered, the number of selected-response (SR) and open-ended (OE) items, and the maximum number of points per form.

		# of Items/Form		Max. # of
Content Area	# of Forms	SR	OE	Points[1]/Form
Reading (RD)	10	35	1	41
Mathematics (MA)	9	40	7	78 - 94[2]
Science (SC)	10	46	8	78 - 82[2]

[1] Summed ratings.

[2] Maximum number of points varied from form to form.

Note that the maximum number of points reflects "summed" ratings for the OE items. Each response to the OE items was scored by two readers, yielding two ratings. The sum of two readings, as opposed to the average of two readings, was used for the calibrations of the items. See Sykes, Yen, & Ito (1996) for results of a comparison between averaging and summing. The OE question in Reading was scored on a 0-3 scale. The rating scale for the OE questions in Mathematics varied from question to question: 0-2, 0-3, 0-4, 0-5 scales were used. In Science, the OE items were rated on a 0-1, 0-2, or 0-3 scale.

Table 1 presents some basic statistics for the forms for the individual samples and the combined samples. The statistics include the case count, the raw-score mean and standard deviation, the Feldt-Raju congeneric reliability coefficient (Qualls, 1995), and the standard error of measurement (SEM). The Feldt-Raju reliability coefficient (F-R) is reported in place of more frequently used measures of internal consistency (e.g., coefficient α, KR-20), because coefficient α would underestimate reliability when a test contains items of multiple formats. The forms were reasonably comparable in reliability. The F-R for the combined group ranged from .90 to .95 for Reading, from .90 to .94 for Mathematics, and from .89 to .94 for Science.

Designs of Equating

The example discussed in the introduction illustrates Approach 1 to the 10-form equating for Reading and Science. Approach 2 involved no "equivalent-groups

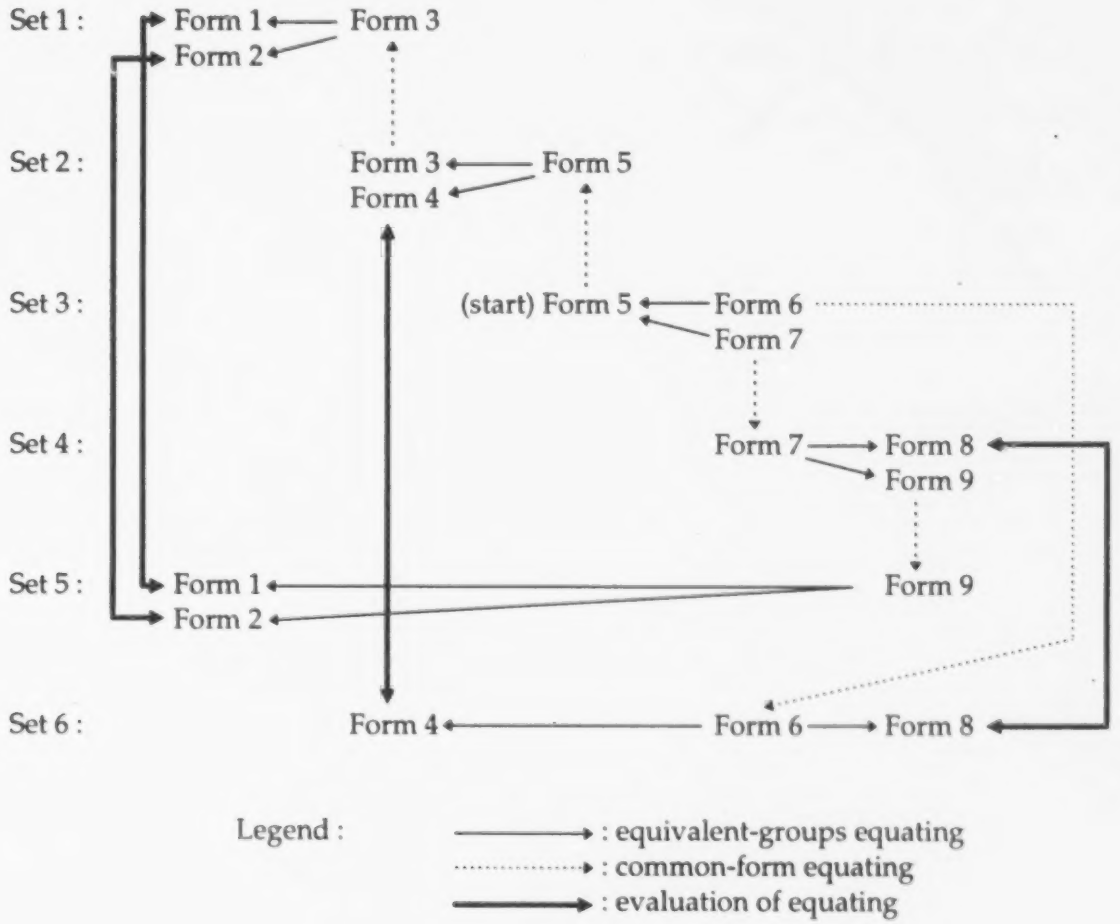
equating". The combined-group calibration in Approach 3 eliminated the need for "common-form equating".

Mathematics involved one fewer form, 9 forms. As with Reading and Science, the nine forms in Mathematics were grouped into six sets.

MA Set	# of Forms	Form
1	3	1, 2, **3**
2	3	**3**, 4, 5
3*	3	5, 6, 7
4	3	7, 8, **9**
5	3	**9**, 1, 2
6	3	4, 6, 8

* Beginning of the links.

The resulting equating design for Mathematics (Approach 1) was very similar to that for Reading and Science.



Procedure

In the first approach ("individual-sample form calibration and explicit equivalent-groups equating"), each form was calibrated twice with the 3-parameter logistic (for SR items) and 2-parameter partial credit (for OE items) models. PARDUX

(Burket, 1991; 1995) was used for all calibrations for this study. PARDUX, similar to BILOG(1983), is a marginal maximum-likelihood estimation program and uses the EM algorithm (Bock & Aitkin, 1981). The program specifies the true ability distribution to have a mean of 0 and standard deviation of 1. The equivalent-groups equatings (thin line) were performed using equipercentile equating. No smoothing was applied in the equipercentile equating. For the common-form equating (broken line), the Stocking and Lord (1983) procedure was used to align test characteristic curves (TCCs). As can be seen in the diagram, the linking was performed in sequence, starting with Set 3 (directional arrows).

With the only exception that forms were not explicitly equated within sets (i.e., no thin lines), Approach 2 ("individual-sample form calibration and no explicit equivalent-groups equating") was identical to Approach 1.

In Approach 3 ("combined-sample form calibration with explicit equivalent-groups equating"), two samples that took the same form were combined and used in the calibration of the form (i.e., no broken lines). The remaining equating procedure was identical to the procedure used in Approach 1.

Once the equatings were completed for all three approaches, the quality of equating was evaluated by comparing the TCCs for the forms that were not used in the equatings and that were at the end or near the end of the connecting linkage (thick line). The forms evaluated are Forms 1, 2, 4, and 8 for all three content areas, and additionally Form 9 for Reading and Science. A total of 14 comparisons were made.

Forms 1 and 2 each were equated with themselves across three sets of forms, and Forms 4, 8, and 9 each were equated back to themselves with one intervening set of forms.

Results

A discrepancy measure proposed by Petersen, Cook, & Stocking (1983) was used to evaluate TCCs. The weighted mean square difference (WMSD) is a sum of the variance of the difference and the squared bias:

$$\text{(WMSD)} = \text{(Variance of difference)} + \text{(Squared bias)}$$

$$\frac{\sum_j f_j d_j^2}{n} = \frac{\sum_j f_j (d_j - \bar{d})^2}{n} + \bar{d}^2,$$

where $d_j = (t_j - t_j)$; t_j is the estimated scaled score for raw score x_j; t_j is the criterion scaled score for x_j; f_j is the frequency of x_j for the estimated form; $n = \sum_j f_j$ for the estimated form; $\bar{d} = \frac{\sum_j f_j d_j}{n}$; and the summation is over that range of x for which extrapolation is unnecessary.

The WMSD has several advantages. First, as an index weighted by the number of cases, "instead of giving equal weight to all values of [raw score], it gives the greatest weight to those values of [raw score] that are most likely to occur" (p.149). Furthermore, the index provides two important pieces of information about the quality of equating: equating bias and standard error of equating. With this index, one can tell

not only which equating method is less biased but also which method provides more stable results.

A cautionary note is necessary about our choice of the criterion form. All the forms were administered simultaneously. Therefore, time was not a factor in determining the criterion form. Moreover, each pair of forms compared (e.g., Form 1 in Set 1 and Form 1 in Set 5) was equi-distant from the starting form of the equating (i.e., Set 3). Because of this characteristic of equal distance from the beginning form, both of the forms in a pair being evaluated were equally suitable for the criterion. Thus, the criterion was decided arbitrarily. To avoid any systematic bias in choosing the criterion form, the criterion form for the pairs was alternated between the smaller-number set and the greater-number set. The criterion is circled in the table below.

Form used in evaluation	Reading		Mathematics		Science	
Form 1	Set 1	(Set 5)	(Set 1)	Set 5	Set 1	(Set 5)
Form 2	(Set 1)	Set 5	Set 1	(Set 5)	(Set 1)	Set 5
Form 4	Set 2	(Set 6)	(Set 2)	Set 6	Set 2	(Set 6)
Form 8	(Set 4)	Set 6	Set 4	(Set 6)	(Set 4)	Set 6
Form 9	Set 4	(Set 6)			Set 4	(Set 6)

Table 2 summarizes the equating results from the 14 comparisons in three statistics: (1) the WSMD or total error, (2) the mean difference or bias, and (3) the standard deviation of difference or standard error.

WMSD or total error in equating

Of the 14 comparisons of the three approaches, Approach 3 exhibited the

smallest WMSD in 12 comparisons. In two thirds of the 12 comparisons where Approach 3 showed superiority, total error for Approach 1 was only somewhat higher than that for Approach 3 (i.e., difference in the WMSD ≤ 50). In all comparisons, however, Approach 2 demonstrated substantially greater total error than either Approach 3 or Approach 1.

Figures 1 through 4 illustrate how relatively small or large values of total error translate to TCCs.

	Content	Form	Approach	Total error
Figure 1	Reading	1	3	18.71
Figure 2	Reading	9	1	89.63
Figure 3	Reading	1	1	389.30
Figure 4	Science	1	2	22764.92

Equating bias

The absolute value of equating bias was the smallest for Approach 3 in eight of the 14 comparisons. For the remaining six comparisons, Approach 1 yielded the smallest absolute mean difference. However, the differences between Approaches 1 and 3 were relatively small in most comparisons. Approach 2 demonstrated considerably greater bias in all but one comparisons.

Although, overall, Approach 3 resulted in the least bias in the majority of the comparisons, this picture did not hold when examined by the content area. Reading was the only content area where Approach 3 consistently had the smallest absolute mean difference. In Mathematics and Science where forms were longer and contained

more OE items, Approach 3 exhibited slightly more bias than Approach 2, respectively, in three out of four comparisons and three out of five comparisons. A content effect has been reported in other studies (e.g., Petersen, Cook, & Stocking, 1983; Holmes & Doody-Bogan, 1983). One possible hypothesis was that having more OE items in Mathematics and Science led to the violation of the unidimensionality assumption and made the results somewhat less conclusive. However, this hypothesis was rejected with confirmatory factor analysis for at least Science.

Standard error of equating

Approach 3 had the smallest standard deviation of the difference in ten of the 14 comparisons. In contrast, Approach 2 showed the greatest instability in all but one comparisons. In six of the ten comparisons favorable to Approach 3, the difference in standard error between Approach 3 and Approach 1 was very small.

Discussion

Overall, Approach 3 ("combined-sample form calibration with explicit equivalent-groups equating") demonstrated the greatest stability, the least bias, and therefore, the least total error. Compared with Approach 3, Approach 1 ("individual-sample form calibration and explicit equivalent-groups equating") showed more error (both bias and standard error), but the differences between these two approaches were generally small. Approach 2 ("individual-sample form calibration and no explicit equivalent-groups equating") exhibited substantially more bias and instability of equating.

The satisfactory results for Approach 3 and Approach 1, relative to those for Approach 2, suggest that spiraling in this case resulted in samples that were not equivalent. As a consequence, the equating results were better with Approaches 3 and 1 that included equivalent-groups equating within sets. In the absence of equivalent samples, as long as forms are not widely different in difficulty, the linking procedure should include the equating of random groups based on the observed-score distributions to improve the quality of linking.

Crouse and Hoover (1991) suggest larger sample size as a means of reducing total error. Sample size probably is a major reason why, compared with Approach 1, Approach 3 resulted in slightly less equating bias and slightly more stability. With regard to sample size, Approach 3 has two advantages over Approach 1. First, Approach 3 estimated 3PL parameters with twice as many cases (approximately 1,000) as Approach 1[1] and therefore had less estimation error contributing to equating error. Second, unlike Approach 1, Approach 3 does not involve the Stocking and Lord 3PL IRT equating method which presumably requires more cases than equipercentile equating. Because of these sample-size-related advantages and also because of the more practical advantage that it involves fewer calibrations, Approach 3 should be preferred to Approach 2 in an equating situation similar to the present one.

As a last note, the number of intervening sets of forms in equating does not seem to have a systematic effect on equating.

---

[1] Despite the smaller sample sizes, Approach 2 produced results that are comparable to, and sometimes even better than, those for Approach 3 in many cases. CTB's internal research, as well as the present results, suggest that the 3PL parameter estimation may be able to be conducted with substantially fewer than 1,000 cases.

References

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

Burket, G.R. (1991, 1995). *PARDUX.* Monterey, CA: CTB McGraw-Hill.

Crouse, J.D., & Hoover, H.D. (1991). *A comparison of the accuracy of equatings resulting from three data collection designs.* Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago.

Holmes, S.E., & Doody-Bogan, E.N. (1983). *An empirical study of vertical equating methods using the three-parameter logistic model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Mislevy, R.J., & Bock, R.D. (1983). *BILOG: Item and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

Qualls, A.L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education, 8,* 111-120.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Sykes, R.C., Yen, W.M., & Ito, K. (1996). *Scaling polytomous items that have been scored by two raters.* A paper to be presented at the annual meeting of the National Council on Measurement in Education, New York.

Table 1
Basic statistics for the forms.

Content	Form	Set	N	Raw Score Mean	SD	F-R	SEM
RD	1	1	521	24.7 (60.2%)	7.5	.943	1.79
		5	479	22.9 (55.9%)	8.7	.958	1.78
		Combined	999	23.9 (58.3%)	8.1	.952	1.77
	2	1	505	23.7 (57.8%)	6.5	.917	1.87
		5	491	21.5 (52.4%)	7.5	.932	1.96
		Combined	995	22.7 (55.4%)	7.1	.928	1.91
	3	1	492	26.6 (64.9%)	6.9	.932	1.80
		2	431	23.1 (56.3%)	8.5	.947	1.96
		Combined	922	25.0 (61.0%)	7.9	.944	1.87
	4	2	407	22.6 (55.1%)	8.3	.930	2.20
		6	549	26.1 (63.7%)	7.4	.925	2.03
		Combined	955	24.6 (60.0%)	8.0	.933	2.07
	5	2	411	18.0 (43.9%)	8.2	.926	2.23
		3	461	20.7 (50.5%)	8.1	.932	2.11
		Combined	871	19.4 (47.3%)	8.3	.931	2.18
	6	3	455	21.5 (52.4%)	8.3	.938	2.07
		6	547	22.5 (54.9%)	8.1	.933	2.10
		Combined	1001	22.0 (53.7%)	8.2	.936	2.07
	7	3	425	19.5 (47.6%)	8.2	.925	2.25
		4	477	19.1 (46.6%)	7.9	.906	2.42
		Combined	901	19.3 (47.1%)	8.1	.915	2.36
	8	4	477	19.0 (46.3%)	7.1	.885	2.41
		6	529	20.6 (50.2%)	7.7	.910	2.31
		Combined	1005	19.8 (48.3%)	7.5	.901	2.36
	9	4	471	23.8 (58.0%)	7.5	.925	2.05
		6	515	25.7 (62.7%)	7.6	.941	1.85
		Combined	985	24.8 (60.5%)	7.6	.935	1.94
	10	4	462	22.6 (55.1%)	8.2	.932	2.14
		5	489	22.2 (54.1%)	7.8	.916	2.26
		Combined	950	22.4 (54.6%)	8.0	.924	2.21
					Mean	SEM	2.08

(continued)

Table 1 (continued)
Basic statistics for the forms.

Content	Form	Set	N	Raw Score Mean	Raw Score SD	F-R	SEM
MA	1	1	580	28.6 (36.7%)	14.7	.944	3.48
		5	444	32.4 (41.5%)	14.9	.938	3.71
		Combined	1023	30.3 (38.8%)	14.9	.942	3.59
	2	1	570	31.7 (34.5%)	18.6	.933	4.81
		5	445	34.1 (37.1%)	17.5	.916	5.07
		Combined	1014	32.7 (35.5%)	18.2	.926	4.95
	3	1	577	36.5 (39.7%)	20.2	.911	6.03
		2	571	35.1(38.2%)	19.9	.908	6.04
		Combined	1147	35.8 (38.9%)	20.1	.909	6.06
	4	2	553	32.4 (36.0%)	19.1	.942	4.60
		6	610	37.3 (41.4%)	19.9	.942	4.79
		Combined	1162	35.0 (38.9%)	19.7	.943	4.70
	5	2	564	31.7 (33.7%)	18.4	.929	4.90
		3	451	35.5 (37.8%)	20.0	.938	4.98
		Combined	1014	33.4 (35.5%)	19.3	.934	4.96
	6	3	444	40.3 (41.5%)	20.8	.930	5.50
		6	601	42.0 (43.3%)	20.1	.922	5.61
		Combined	1044	41.3 (42.6%)	20.5	.926	5.58
	7	3	452	37.3 (36.9%)	22.0	.958	4.51
		4	626	38.2 (37.8%)	18.2	.925	4.98
		Combined	1077	37.8 (37.4%)	19.8	.942	4.77
	8	4	610	30.8 (30.5%)	15.2	.902	4.76
		6	610	31.2 (30.9%)	17.3	.925	4.74
		Combined	1219	31.0 (30.7%)	16.2	.915	4.72
	9	4	617	32.6 (37.9%)	15.4	.895	4.99
		5	447	32.6 (37.9%)	15.0	.900	4.74
		Combined	1063	32.6 (37.9%)	15.2	.896	4.90
					Mean	SEM	4.91

(continued)

Table 1 (continued)
Basic statistics for the forms.

Content	Form	Set	N	Raw Score Mean	SD	F-R	SEM
SC	1	1	514	36.8 (46.0%)	15.7	.929	4.18
		5	521	41.3 (51.6%)	14.3	.913	4.22
		Combined	1034	39.1 (48.9%)	15.2	.923	4.22
	2	1	507	33.6 (42.0%)	14.4	.917	4.15
		5	541	37.4 (46.8%)	14.1	.913	4.16
		Combined	1047	35.6 (44.5%)	14.4	.917	4.15
	3	1	507	32.8 (40.0%)	14.9	.908	4.52
		2	538	32.6 (39.8%)	15.0	.910	4.50
		Combined	1044	32.7 (39.9%)	15.0	.909	4.52
	4	2	519	35.3 (44.1%)	14.4	.921	4.05
		6	544	34.1 (42.6%)	14.3	.921	4.02
		Combined	1062	34.7 (43.4%)	14.4	.921	4.05
	5	2	526	38.6 (49.5%)	15.6	.946	3.63
		3	477	39.2 (50.3%)	14.0	.930	3.70
		Combined	1002	38.9 (49.9%)	14.9	.939	3.68
	6	3	471	34.3 (42.9%)	11.8	.888	3.95
		6	545	33.9 (42.4%)	13.4	.914	3.93
		Combined	1015	34.1 (42.6%)	12.7	.903	3.96
	7	3	469	37.4 (45.6%)	14.2	.895	4.60
		4	404	36.8 (44.9%)	13.8	.890	4.58
		Combined	872	37.1 (45.2%)	14.0	.893	4.58
	8	4	395	37.1 (46.4%)	15.1	.931	3.97
		6	549	35.9 (44.9%)	15.3	.934	3.93
		Combined	943	36.4 (45.5%)	15.3	.932	3.99
	9	4	405	35.9 (44.9%)	15.6	.931	4.10
		6	549	36.1 (45.1%)	16.8	.946	3.90
		Combined	953	36.0 (45.0%)	16.3	.940	3.99
	10	4	403	36.3 (46.5%)	13.9	.908	4.22
		5	531	37.7 (48.3%)	14.1	.909	4.25
		Combined	933	37.1 (47.6%)	14.0	.908	4.25
					Mean SEM		4.13

[1] Number of standard errors of difference between two means (SEDs) in the mean-RS difference for a given form in two different sets.

Table 2
Equating bias, standard error, and total error

		Reading		
Form	Error	App. 1	App. 2	App. 3
	WMSD	389.30	7109.01	18.71
1	Bias	-2.86	82.51	1.32
	SD of Difference	19.52	17.33	4.12
	WMSD	419.38	6671.55	23.67
2	Bias	9.54	-78.69	1.66
	SD of Difference	18.12	21.89	4.57
	WMSD	44.36	2438.90	29.05
4	Bias	1.05	48.24	-0.93
	SD of Difference	6.58	10.58	5.31
	WMSD	128.09	1335.77	87.66
8	Bias	-8.88	35.78	-3.36
	SD of Difference	7.01	7.46	8.74
	WMSD	89.63	1329.84	69.97
9	Bias	9.27	-35.86	7.51
	SD of Difference	1.95	6.65	3.69

(continued)

Table 2 (continued)
Equating bias, standard error, and total error

		Mathematics		
Form	Error	App. 1	App. 2	App. 3
	WMSD	54.96	416.30	31.96
1	Bias	2.59	16.40	1.74
	SD of Difference	6.95	12.14	5.38
	WMSD	97.76	252.13	48.36
2	Bias	5.91	8.88	6.08
	SD of Difference	7.93	13.16	3.38
	WMSD	13.84	13325.14	34.05
4	Bias	-0.50	-109.48	-2.64
	SD of Difference	3.69	36.59	5.20
	WMSD	35.79	14403.94	31.84
8	Bias	1.82	114.45	2.65
	SD of Difference	5.70	36.13	4.98

(continued)

Table 2 (continued)
Equating bias, standard error, and total error

Science				
Form	Error	App. 1	App. 2	App. 3
1	WMSD	116.24	22764.92	105.95
	Bias	-10.43	-150.62	-9.84
	SD of Difference	2.72	8.86	3.02
2	WMSD	138.98	22522.36	130.94
	Bias	10.77	148.09	11.30
	SD of Difference	4.79	24.32	1.83
4	WMSD	69.71	1269.30	108.93
	Bias	6.35	-33.06	9.47
	SD of Difference	5.42	13.28	4.40
8	WMSD	181.01	7991.25	62.66
	Bias	-6.32	-88.32	-4.80
	SD of Difference	11.88	13.80	6.29
9	WMSD	158.78	8644.16	8.88
	Bias	0.69	90.46	2.74
	SD of Difference	12.58	21.46	1.17

Figure 1. TCC's for Reading Form 1,
Equating Approach 3 (WMSD = 18.71)

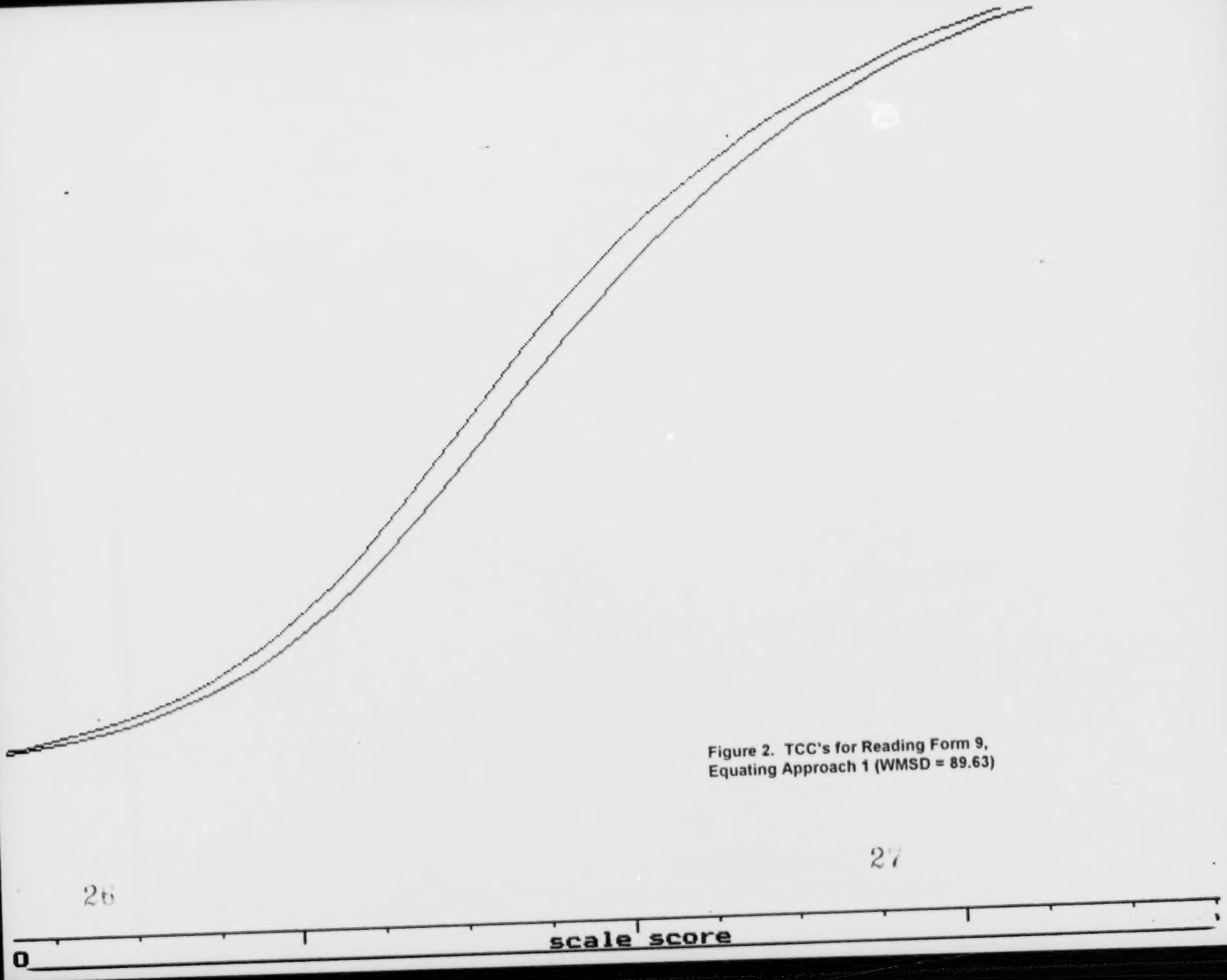


Figure 2. TCC's for Reading Form 9,
Equating Approach 1 (WMSD = 89.63)

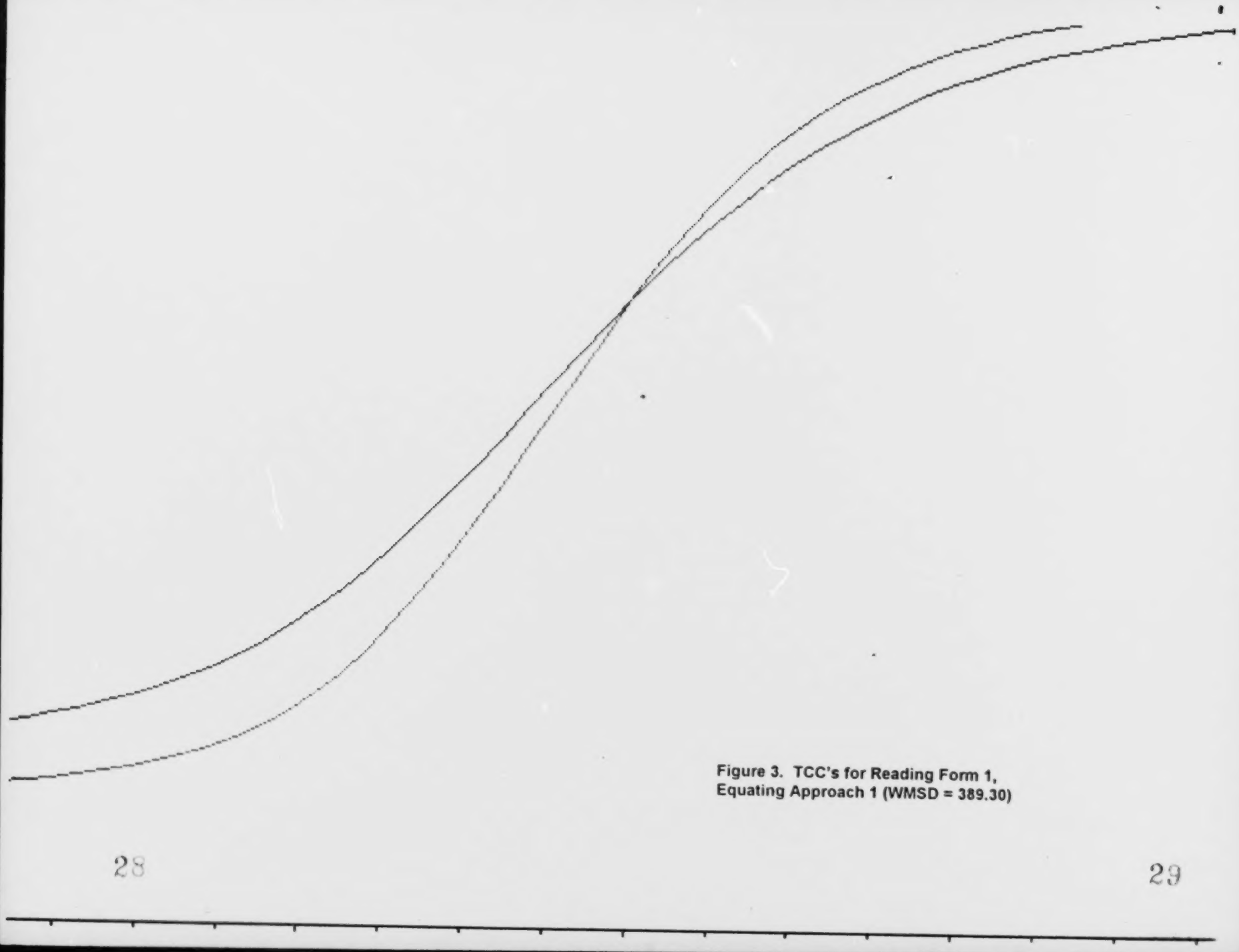


Figure 3. TCC's for Reading Form 1,
Equating Approach 1 (WMSD = 389.30)

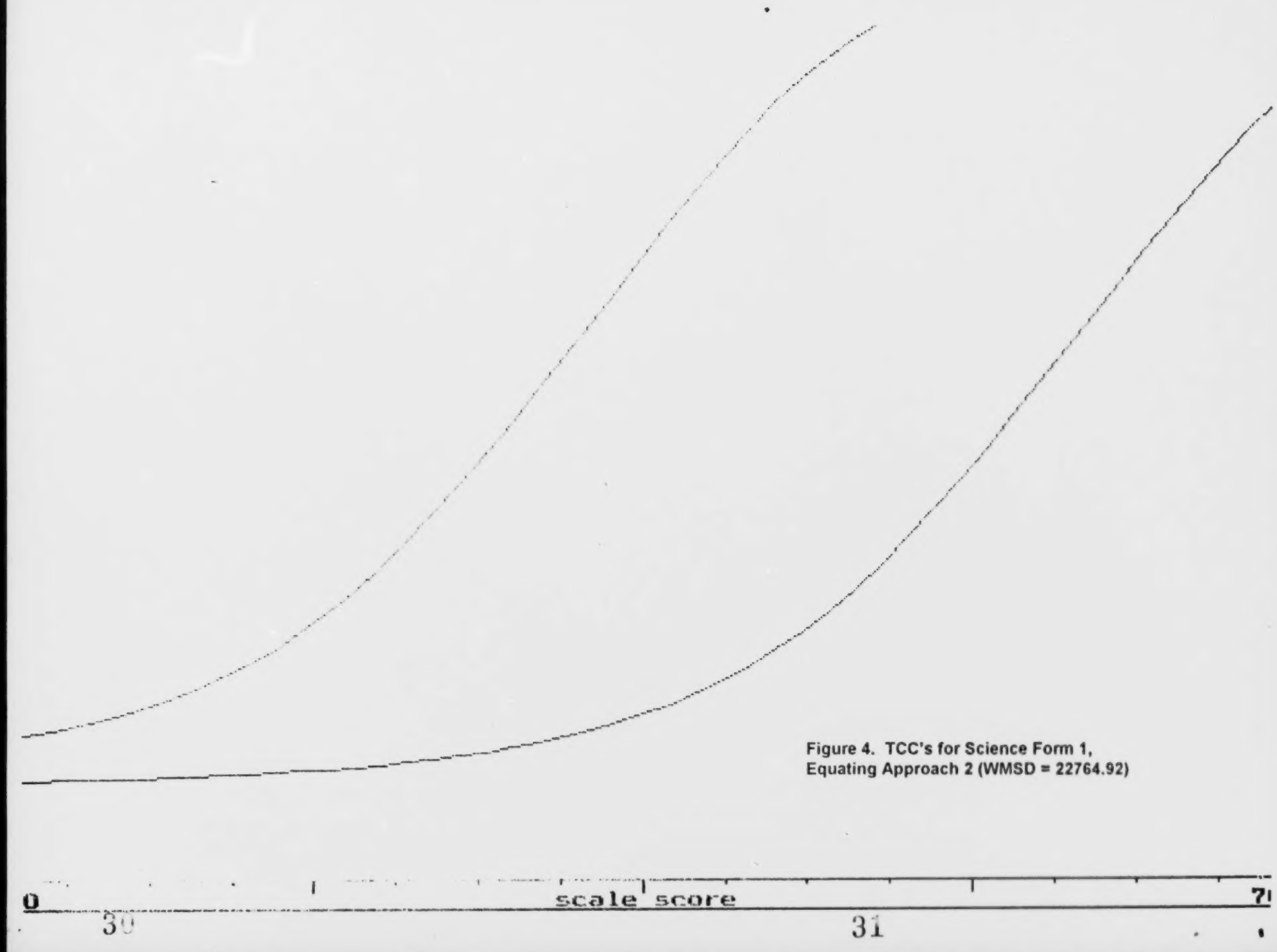


Figure 4. TCC's for Science Form 1,
Equating Approach 2 (WMSD = 22764.92)